Data 6, Summer 2023 **Weichert and Ferrari**

Algorithmic Bias Solutions

Discussion 3 *Summer 2023*

A common misconception about computers and computer algorithms[1] is that since computers aren't human, they must act without bias. After all a computer or an algorithm only does what it's told, right?

In Data 6, we've already seen how data and computing have massive impacts on the 'real world', especially in driving political or policy decisions. In this discussion, we will introduce you to the concept and study of algorithmic bias. Specifically, you will read and see examples of how algorithms *embed* or *perpetuate* certain biases, assumptions and misrepresentations that are either made by the programmer, or present in the data (or lack of data).

1 Are Algorithms Fair?

1. Start by reading "Why algorithms can be racist and sexist" by Rebecca Heilweil. This article provides an overview of the many ways in which algorithms may encode biases against certain groups of people. List at least two examples of biased algorithms mentioned in the article and explain how they are biased.

> **Solution:** Examples include targeting political ads, job application screening, predicting fire risk, determining police deployment, and facial recognition.

2. The article also includes quotes from Nicol Turner-Lee, who studies algorithmic bias at the Brookings Institution. Turner-Lee says there are two primary ways to think about algorithmic bias. What are they and how are they different?

> **Solution:** Turner-Lee talks about two aspects of algorithmic bias: accuracy and impact. Accuracy refers to the 'correctness' of an algorithm across different populations or demographics. Impact is perhaps a more subjective measure, referring to how the decisions that an algorithm makes may influence the lives of different people.

[1] An algorithm is a set of instructions for solving a problem, like sorting a deck of cards or deciding whether to approve an applicant for a credit card or loan.

3. One way that algorithms, especially artificial intelligence or machine learning algorithms, 'become' biased is by making decisions based on *training data* [2] that is not fully representative of the population. So when an algorithm is given 'bad' data, it will tend to make 'bad' decisions. Does that mean that if an algorithm is given data that is representative of the population, then the algorithm will necessarily make 'fair' decisions?

> **Solution:** No. Even with representative training data, algorithms are still prone to bias because the process of creating or coding that algorithm can embed the biases of the programmer. In the article Turner-Lee talks about the need to "think about who gets a seat at the table when these systems are proposed, since those people ultimately shape the discussion about ethical deployments of their technology."

4. What are some ways to hold algorithms and programmers accountable? Do you have thoughts on algorithmic bias and fairness?

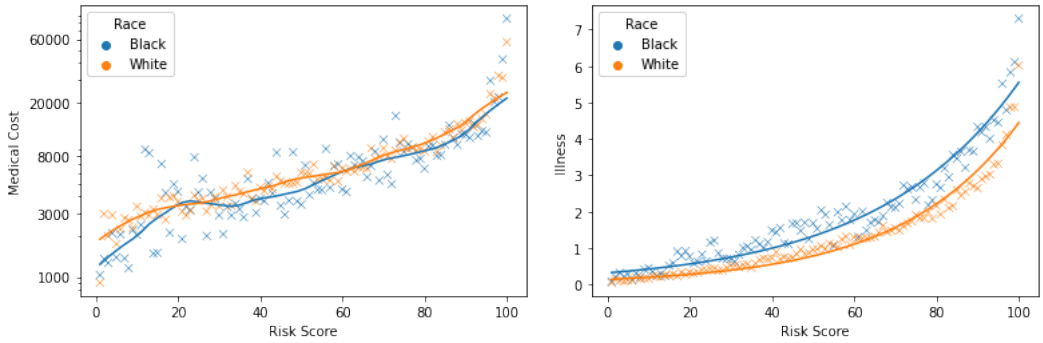
> **Solution:** As noted in the article, a good but not necessarily sufficient first step is to build in more transparency into how algorithms are making their decisions. Being able to see how an algorithm came to a decision will make it easy to hold the algorithm (and its programmers) accountable for bias and unfair decision-making. But Heilweil also notes that the laws and accountability mechanisms for biased decision-making (e.g. for employment) need to be updated to include algorithms.

2 Bias in Health Risk Algorithms

This part of the discussion is based on the "ML Failures lab: Dissecting Racial Bias in a Medical Risk Score" by Nick Merrill, Inderpal Kaur, and Samuel Greenberg.

This lab examines an artificial intelligence algorithm that predicts patient "health risk scores" that are widely used in the medical profession. The algorithm is given information about patients' medical histories and how much patients spend on health care (called *medical cost*). Using this data, the algorithm assigns each patient a risk score that is supposed to tell doctors how likely a patient is to suffer from certain illnesses. These risk scores are used by doctors and hospitals to prioritize certain patients over others. If you're interested in learning more, we recommend reading this research paper by Obermeyer et al.

[2] Training data is the data used to show an algorithm how to form connections between inputs and outputs. An algorithm uses the training data to 'learn' examples of how inputs relate to outputs.



(a) Risk Score and Medical Cost (b) Risk Score and Chronic Illnesses

Figure 1: Health Risk Algorithm Predictions for White and Black Patients

1. Figure 1(a) shows the risk score calculated by the algorithm for patients with different medical costs. What should we expect to see if the algorithm assigns risk scores irrespective of race?

> **Solution:** If race does not play a role in the algorithm's decision-making, the trend lines for risk score vs. medical cost should be identical for both white and Black patients. With this algorithm, however, that is not the case.

2. Figure 1(b) shows predicted risk scores for patients based on the number of chronic illnesses. What can this chart tell us about the fairness of the risk algorithm?

> **Solution:** The second figure shows more clearly the bias in the algorithm. For all risk scores, the algorithm tends to assign *lower* risk scores to Black patients with a certain number of chronic illnesses than it would for white patients. You can see this by drawing a horizontal line at any point and see where that line intersects the blue and orange trend lines. If risk scores are used to prioritize the provision of health care resources, this algorithm may result in more white patients being prioritized over Black patients with the same number of chronic illnesses.

3. If the algorithm wasn't given a patient's race as an input, would that eliminate all bias from the process of predicting risk scores?

> **Solution:** No. In the case of an algorithm to predict 'risk scores', that algorithm is likely to reproduce the biases that are present in the training data (i.e. risk score examples). So if there is an underlying bias in the risk scores that leads to doctors to generally assign a lower risk score to a Black patient than they would for a similar

white patient, that bias will likely be translated into the risk score algorithm. We must not assume that the idea of a medical risk score itself is without bias. In fact, studies (like this one) have documented how doctors tend to downplay the pain of Black patients compared to white patients.